Apprendre l'apprentissage automatique : un retour d'expérience

Noëlie Debs, Sergio Peignier, Clément Douarre, Théo Jourdan, Christophe Rigotti, Carole Frindel

▶ **To cite this version:**

Noëlie Debs, Sergio Peignier, Clément Douarre, Théo Jourdan, Christophe Rigotti, et al.. Apprendre l'apprentissage automatique : un retour d'expérience. CETSIS 2021 - Colloque de l'Enseignement des Technologies et des Sciences de l'Information et des Systèmes, Jun 2021, Valenciennes, France. pp.1-5. hal-03341954

HAL Id: hal-03341954
https://hal.science/hal-03341954v1

Submitted on 13 Sep 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Apprendre l'apprentissage automatique : un retour d'expérience

Noëlie Debs[1,2], Sergio Peignier[1,3], Clément Douarre[1,4], Théo Jourdan[1,2], Christophe Rigotti[1,4], Carole Frindel[1,2]
carole.frindel@insa-lyon.fr
1 Département Biosciences, INSA de Lyon ;
2 : Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69621, LYON, France
3 : Univ Lyon, INSA-Lyon, CNRS, INRA, BF2I UMR0203, F-69621, Villeurbanne, France
4 : Univ Lyon, INSA-Lyon, CNRS, INRIA, LIRIS, UMR5205, F-69621 Villeurbanne, France

RESUME : Dans cet article, nous présentons un retour d'expérience sur un module d'initiation à l'apprentissage automatique de 3 crédits ECTS que nous avons créé pour des élèves ingénieurs de l'INSA de Lyon. Nous présentons la structuration de ce module ainsi que les spécificités pédagogiques liées au faible nombre d'heures de face à face au regard de la complexité du sujet. Nous montrons des exemples de cas d'usage proposés aux étudiants et faisons état de leurs réactions. L'ensemble est complété par des liens bibliographiques vers les sites des outils numériques libres et références pédagogiques utilisées.

Mots clés : apprentissage automatique, modélisation, prédiction, retour d'expérience.

1. INTRODUCTION

Nous sommes entrés dans un monde où les données sont en passe de devenir l'essence même de la connaissance et de l'information. Il n'est donc pas étonnant que l'apprentissage automatique soit devenu une compétence de base pour un large panel de formations d'ingénieurs et que le métier de « data scientist » soit très prisé par les entreprises [1,2,3]. Bien que l'apprentissage automatique soit un domaine de l'informatique, il diffère des approches informatiques traditionnelles. En effet, les algorithmes d'apprentissage automatique sont constitués d'un ensemble d'instructions explicitement programmées pour donner à l'ordinateur la capacité d'« apprendre » à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâche.
Pour autant, l'apprentissage automatique est une discipline relativement avancée qui nécessite des bases à la fois en statistiques, pour comprendre la modélisation à partir d'un échantillon de données, mais aussi en mathématiques et notamment en algèbre, pour réaliser des opérations sur des jeux de données, ainsi qu'en programmation, pour mettre en œuvre l'ensemble du processus d'apprentissage automatique et comprendre les bibliothèques mises à disposition dans différents langages de programmation. Les enseignements en apprentissage automatique se positionnent par conséquent souvent à un niveau L3 ou au-delà. Ces nombreux prérequis nécessaires ont été contournés en créant des cours plus appliqués, où l'accent est davantage mis sur les compétences pratiques, telles que la mise en place d'outils, la compréhension des moyens de visualisation et la préparation de données. Bien qu'il s'agisse de compétences « bas niveau » nécessaires pour appliquer l'apprentissage automatique, des idées fausses peuvent se cacher sous ces compétences, ce qui peut conduire les étudiants à croire que les configurations par défaut des algorithmes d'apprentissage automatique sont adaptées à tous les types de données. Ainsi un classificateur qui a été construit avec succès serait considéré « correct » comme les programmes informatiques déterministes sont corrects, malgré le large éventail de failles possibles liées aux données.

Maintenant que l'apprentissage automatique atteint les masses plus larges par le biais des cours d'enseignement diffusés sur internet (MOOC) [4,5,6], d'outils qui se revendiquent profitables sans compétences techniques [7,8,9] et même indirectement par le biais des produits de consommation [10,11], nous craignons que ces mêmes idées fausses existent, mais à une échelle beaucoup plus grande. Nous connaissons encore peu de choses sur ce que les élèves doivent savoir, comment l'enseigner et quelles connaissances les enseignants doivent avoir pour réussir à l'enseigner. C'est, dans ce contexte, que se situe notre contribution. Afin de proposer une réponse à ce défi pédagogique, nous tentons par le biais de ce module de découvrir les connaissances en contenu pédagogique nécessaires à l'enseignement des concepts en apprentissage automatique. Pour ce faire, nous avons tâché d'identifier des représentations utiles pour les concepts de l'apprentissage automatique, des cas d'usage afin d'illustrer l'apprentissage automatique et apprendre à connaître les concepts difficiles de l'apprentissage automatique ainsi que les erreurs courantes commises par les apprenants lors de l'application de l'apprentissage automatique.

Dans le cadre de cet article, nous présentons un module d'initiation à l'apprentissage automatique de 40 heures [12] que nous avons créé à destination d'étudiants ingénieurs : les étudiants de 4ème année de la filière de Bio-Informatique et Modélisation (BIM) de l'institut national des sciences appliquées (INSA) de Lyon [13]. Dans la suite, nous détaillons le profil des étudiants puis présentons la structuration et le contenu en donnant au lecteur des liens vers des ressources pédagogiques utiles pour monter ce type d'enseignement. Il est à noter que nous avons privilégié des outils logiciels libres et fait le lien avec des communautés actives et mondiales de développeurs. Compte tenu du faible volume horaire et la

complexité du sujet en jeu, nous avons développé une pédagogie par projet pour l'évaluation. Nous donnons des exemples de réalisations et relatons les réactions des étudiants.

2. PROFIL DES ETUDIANTS

La filière Bio-Informatique et Modélisation de l'INSA de Lyon est une filière de deuxième cycle du département Biosciences et comprend environ 25 élèves par promotion. Elle vise à former en 3 ans des ingénieurs à l'interface entre la biologie, les mathématiques et l'informatique afin de répondre à la demande croissante émanant de l'augmentation considérable de la masse de données biologiques issues des techniques dites à haut débit. Dans ce cadre, l'apprentissage automatique prend tout son sens, d'où la nécessité de créer rapidement un module dédié à cette discipline afin de former les étudiants à l'intelligence artificielle et répondre à la demande grandissante du tissu économique local, national et internationale en « data scientist ».

Biologistes, ces ingénieurs doivent être capables d'analyser, de traiter des données biologiques et d'en extraire les informations pertinentes. Algorithmiciens, ils sont à même d'élaborer des outils informatiques pour analyser ces informations, afin d'émettre des hypothèses, et, à partir de ces dernières, de créer des modèles mimant les systèmes biologiques afin de mieux comprendre les processus du vivant.
Lors de la 4ème année, le module Intelligence Artificielle (IA) leur est proposé. L'idée de ce module est d'illustrer le champ d'application de l'intelligence artificielle dans le domaine des sciences de la vie ainsi que les besoins émergents en « Data Science » impliquant l'apprentissage automatique et d'autres approches d'intelligence artificielle pour répondre à des questions de la recherche en biologie.

3. STRUCTURATION DU MODULE

Le module a été lancé en 2018/2019 et comprend 40 heures qui s'organisent selon l'enchaînement détaillé dans la Figure 1. Ce module a pour but l'acquisition des compétences nécessaires à l'identification et la réalisation des spécifications techniques d'une application intégrant de l'intelligence artificielle : gestion des données, fonctionnalités, algorithmes, etc. En maîtrisant ces compétences, les étudiants seront en capacité d'aller vers les domaines du machine learning ou du deep learning. Chaque partie (flèche bleue sur la Figure 1) s'organise sous la forme de travaux dirigés où les méthodes abordées sont données sur la droite. Les thèmes abordés dans ces parties comprennent 1) les algorithmes de base d'apprentissage non supervisé (clustering) et 2) supervisé (classification), 3) les différentes grandes méthodes d'optimisation et 4) une introduction au deep learning. Ces cours s'appuient sur des modules préexistants au sein de la filière, qui ont été expérimentés et ont évolué au cours des 10 dernières années.

Cette trame plutôt classique comprend un certain nombre d'originalités listées ci-après afin de sélectionner les outils et bibliothèques logicielles adaptés et permettre aux étudiants de manipuler les concepts sur des vraies données et des vrais cas d'usage en recherche.
Enfin, en termes d'évaluation du travail des étudiants, un compte-rendu de TD est demandé pour les parties « Classification » et « Optimisation » auquel s'ajoute une note de projet (détaillé dans la Section 5).

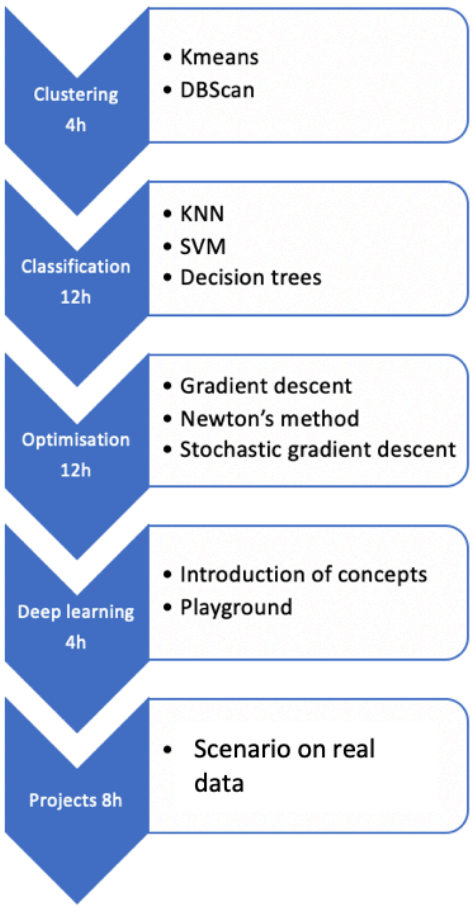

Figure 1: Enchaînement détaillé du module. KNN : K-Nearest Neighbors, SVM : Support Vector Machine.

4. OUTILS : R OU PYTHON ?

De vifs débats existent pour déterminer lequel de ces deux langages il faut utiliser par soucis d'efficacité, et bien sûr surtout pour des raisons pratiques [14].
Selon [15] et notre expérience, tous les frameworks existants et les nouveaux algorithmes d'apprentissage automatiques sont implémentés en Python. La communauté est aussi beaucoup plus importante et surtout Python bénéficie d'un énorme écosystème. Par contre R, qui est un langage traditionnellement utilisé par des statisticiens, incorpore davantage de méthodes d'analyse

statistique à visée exploratoire et peut donc s'avérer plus efficace sur le terrain.
Ce cours étant orienté principalement sur l'apprentissage automatique et s'intégrant à une optique métier, nous avons pris le parti d'utiliser le langage Python. Nous présentons par la suite quelques outils de l'écosystème Python que nous jugeons très intéressants dans le cadre de la mise en œuvre de l'apprentissage automatique.

Figure 2: Le langage Python et son écosystème adapté à l'apprentissage automatique. Schéma issu de http://joseph-salmon.eu/HMMA238.html

4.1 Jupyter Notebook

Comme environnement d'exécution nous avons retenu Jupyter Notebook [16], qui est une application web permettant de stocker des lignes de code Python, les résultats de l'exécution de ces dernières (graphiques, tableaux, etc.) et du texte formaté. Cela est particulièrement adapté à l'apprentissage automatique qui est une discipline par essence itérative : il faut souvent tenter plusieurs approches et étudier les résultats avant de décider de la bonne façon de traiter un problème.

4.2 Les librairies Python pour l'apprentissage automatique

Python possède un ensemble robuste de librairies qui permettent aux étudiants de mettre en œuvre facilement des méthodes d'apprentissage automatique sans avoir à réécrire de nombreuses lignes de code (voir Figure 2). Dans le cadre de ce module nous utilisons NumPy et SciPy pour les calculs, Matplotlib et Seaborn pour la visualisation, Scikit-learn [17] pour les algorithmes d'apprentissage automatique, Pandas pour la gestion des données et Tensorflow, Pytorch et Keras pour le deep learning.

4.3 Outil ludique pour l'introduction au deep learning

Le deep learning est un ensemble de méthodes d'apprentissage automatique tentant de modéliser avec un haut niveau d'abstraction des données grâce à des architectures complexes de différentes transformations non linéaires. Ces techniques ont permis des progrès importants et rapides dans les domaines de l'analyse des signaux et des images mais cependant peuvent être difficiles à mettre en œuvre en enseignement étant donné les infrastructures matérielles nécessaires (ordinateurs puissantes, cartes graphiques, etc). Une bonne alternative consiste à utiliser TensorFlow Playground [18] qui est une visualisation interactive dans le navigateur Web de réseaux de neurones. Il contient une petite bibliothèque de réseaux neuronaux et plusieurs cas d'application simples, ainsi que des visualisations didactiques permettant de comprendre plus facilement comment fonctionne un tel réseau. Les étudiants peuvent ainsi simuler, en temps réel de petits réseaux de neurones et voir les résultats, le tout dans leur navigateur sans besoins matériels importants.
Pour l'avenir proche, nous envisageons également de tester Google Colab qui est un outil complet pour entraîner et tester rapidement des modèles d'apprentissage automatique sans avoir de contrainte matérielle. En effet, Google y met à disposition, gratuitement pour le cadre académique, des central processing units (CPU), graphics processing units (GPU) et tensor processing units (TPU).

5. APPRENTISSAGE PAR PROJET

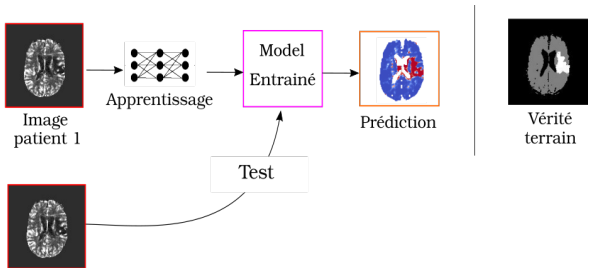

Figure 3: Aperçu de la chaine d'expériences du TP sur images médicales. Les élèves ont à disposition les images IRM de 2 patients atteint d'AVC. Ils apprennent un modèle supervisé (machine à vecteur de support) à partir des images d'un patient, et testent ce modèle sur les images d'un autre patient. Leur modèle est en mesure à la sortie de prédire une carte de probabilité d'infarctus (en bleu les tissus sains, en rouge les tissus lésionnels). Cette carte prédite est à comparer avec une vérité terrain : le masque de la vraie lésion finale.

Le module se conclut par un projet appliqué à des jeux de données réelles issus dans nos problématiques de recherche. Ce projet s'étalant sur 8h permet d'aborder sereinement la problématique, l'application visée, la campagne de collecte de données et l'approche d'apprentissage automatique qui a été appliquée. Puis les étudiants sont invités à travailler en groupes de projet pour reproduire l'expérience de bout en bout. L'idée étant de leur faire identifier les possibilités de l'apprentissage automatique et surtout de leur en faire appréhender les limites.

Dans ce contexte, les projets déjà proposés font écho aux sujets de thèse de deux doctorants rattachés à l'enseignement de ce module [20,21]. Tandis que l'un s'intéresse aux données de type « images », l'autre se concentre sur des données de type « signaux » issues de capteurs sans fil (accéléromètre et gyroscope). Les deux sujets sont en lien avec la médecine (rappelons que la filière a un tronc

commun avec une forte composante en biologie). Ainsi le premier [20] vise à prédire la forme et l'étendue finale de la lésion développée lors d'une accident vasculaire cérébral (AVC) à partir d'une imagerie par résonance magnétique (IRM) mis en place lors de l'arrivée du patient à l'hôpital (voir Figure 3). L'autre sujet [21], également en lien avec l'AVC, cherche à détecter le niveau d'activité d'un patient lors de sa rééducation pour connaître son évolution et adapter les soins et le suivi durant cette phase qui se déroule à son domicile (voir Figure 4).

Il est intéressant de mentionner que le deuxième sujet fait écho à une matière dispensée au sein de la filière qu'est l'éthique [22]. En effet, autour de la problématique des capteurs sans fil nous sensibilisons également les étudiants au contenu des données et de ce qu'elles dévoilent de notre vie privée [23]. L'idée est de les confronter à une problématique éthique concrète et d'y apporter des solutions techniques en les faisant réfléchir à comment mettre en œuvre un apprentissage automatique qui veille à établir un compromis entre l'optimisation de l'utilité des données et en parallèle leur anonymisation [24].

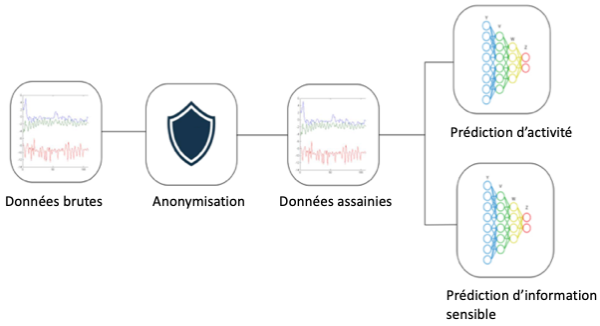

Figure 4: Aperçu de la chaine d'expériences du TP sur les signaux issus de capteurs sans fil. Les élèves ont à disposition les signaux (accéléromètre et gyroscope) de 24 patients avec une annotation de 2 activités statiques (assis et allongé) et 4 activités dynamiques (marche, montée et descente d'escaliers, course). Ils apprennent un modèle supervisé (forêt aléatoire) à partir des signaux sur 2 tâches distinctes : la reconnaissance de l'activité et de l'identité du patient. En analysant l'importance accordé à chaque descripteur pour chaque des 2 tâches, les étudiants doivent trouver un moyen de gommer les descripteurs nécessaires à la tâche d'identification.

6. EVALUATION DU MODULE PAR LES ETUDIENTS

Un questionnaire a été créé sur le site AskaBox [25] afin de permettre aux étudiants d'évaluer le module. L'idée était de vérifier leur satisfaction sur cette nouvelle proposition de module et de prendre en compte d'éventuelles remarques ou critiques afin de l'améliorer pour sa prochaine édition. Première observation, le taux de réponse a été très élevé : 22 étudiants sur 25 ont participé au sondage avec un taux de questionnaires intégralement documentés de 90%. Premier fait marquant, les étudiants pensent que l'apprentissage automatique est une discipline importante pour leur futur métier d'ingénieur (voir Figure 5). Par ailleurs, ils sont globalement satisfaits de la mise en œuvre de cet enseignement (voir Figure 6) et enfin pensent avoir assimiler le contenu de l'enseignement (voir Figure 7).

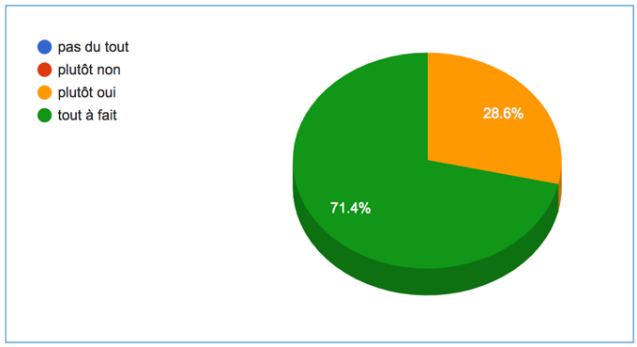

Figure 5: Analyse des réponses des étudiants à l'affirmation « Les objectifs de la discipline Intelligence Artificielle me semblent intéressants pour ma formation de futur ingénieur. »

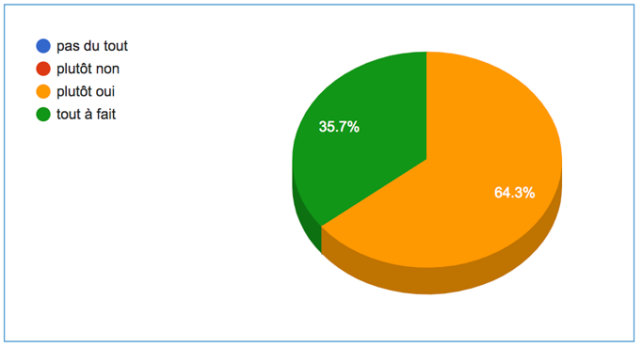

Figure 6: Analyse des réponses des étudiants à l'affirmation « La formation est en adéquation avec les acquis de l'apprentissage visés par cet enseignement (connaissances, capacités et compétences). »

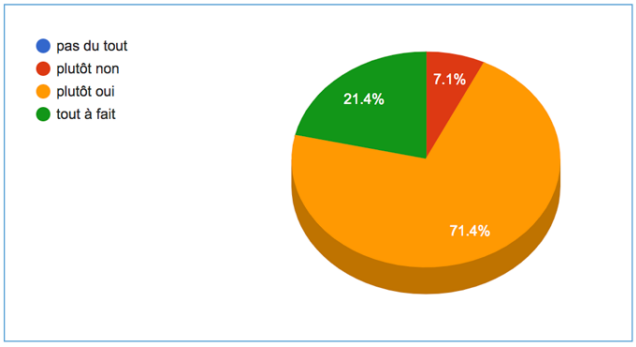

Figure 7: Analyse des réponses des étudiants à l'affirmation « J'ai l'impression d'avoir assimilé le contenu de la discipline et de savoir le mettre en pratique. »

Le questionnaire comprenait également deux champs libres : le premier sur les aspects appréciés dans le cadre

du module et le deuxième pour transmettre des remarques ou critiques à propos du module. Concernant les points positifs, deux aspects ressortent des réponses des étudiants : les outils numériques qui sont facilement accessibles et implantables sur leurs machines personnelles et les applications sur des cas concrets. Concernant les points négatifs, un aspect ressort notablement : le niveau d'investissement réclamé par le module notamment dû au faible nombre d'heures consacré au vu de la quantité de nouveaux concepts à intégrer.

7. CONCLUSION

Nous avons présenté la philosophie, la structuration et le contenu d'un module relativement court d'apprentissage automatique à vocation pratique et expérimental. Dans ce module, nous avons fait le choix de privilégier des cadres de mathématiques discrètes pour aller vers des présentations algorithmiques proches de la culture de nos étudiants (voir Figures 2 et 3). Nous avons donné dans cet article des exemples de réalisations de nos étudiants. Nous jugeons certaines de ces réalisations spectaculaires au regard du faible nombre d'heures allouées à notre module. Ceci s'explique sans doute par le fait que nous avons pu nous appuyer sur les solides compétences en informatique et en statistiques de nos étudiants et surtout par l'appui de jeunes doctorants passionnés par la transmission de leurs connaissances en lien avec leur sujet de thèse. Les exemples de cas d'usage que nous avons montrés pourraient toutefois être adaptés et notamment servir à diversifier les exemples d'illustrations pour des cours d'apprentissage donnés dans le cadre de la finance, du génie industriel, etc.

Bibliographie

[1] LeMagIT, article du 9 janvier 2020, https://www.lemagit.fr/actualites/252476527/Data-scientist-la-demande-reste-tres-forte-mais-evolue-rapidement, consulté le 6 mai 2021.

[2] Le journal du Net, article du 15 avril 2019, https://www.journaldunet.com/management/formation/1423287-quelles-sont-les-competences-en-ia-les-plus-prisees-en-france/, consulté le 6 mai 2021

[3] Widoobiz, article du 23 anvier 2020, https://www.widoobiz.com/2020/01/23/etre-developpeur-finalement-cest-savoir-dialoguer-avec-une-machine-aude-barral-cofondatrice-de-codingame/, consulté le 6 mai 2021

[4] Coursera, mood sur l'apprentissage automatique, https://www.coursera.org/learn/machine-learning, consulté le 6 mai 2021

[5] Mooc Francophone, « Initiez-vous au machine learning », https://mooc-francophone.com/cours/initiez-vous-au-machine-learning/, consulté le 6 mai 2021

[6] OpenClassRooms, « Initiez-vous au machine learning », https://openclassrooms.com/fr/courses/4011851-initiez-vous-au-machine-learning, consulté le 6 mai 2021

[7] BigML, https://bigml.com/, consulté le 6 mai 2021

[8] Progress, https://www.progress.com/datarpm, consulté le 6 mai 2021

[9] AutoKeras, https://autokeras.com/, consulté le 6 mai 2021

[10] Forbes, article du 30 août 2018, https://www.forbes.com/sites/bernardmarr/2018/04/30/27-incredible-examples-of-ai-and-machine-learning-in-practice/#326c75137502, consulté le 6 mai 2021

[11] WordStream, article du 12 août 2019 https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications, consulté le 6 mai 2021

[12] Fiche ECTS du cours d'IA de l'INSA Lyon, http://planete.insa-lyon.fr/scolpeda/f/ects?id=38172&_lang=fr, consulté le 6 mai 2021

[13] Site du département Biosciences de l'INSA Lyon, https://biosciences.insa-lyon.fr/, consulté le 6 mai 2021

[14] OpenSource, article du 22 novembre 2016, https://opensource.com/article/16/11/python-vs-r-machine-learning-data-analysis, consulté le 6 mai 2021

[15] Ozgur, C., Colliau, T., Rogers, G., Hughes, Z., & Myer-Tyson, B. (2017). *MatLab vs. Python vs. R. Journal of Data Science, 15(3), 355-372.*

[16] Jupyter, https://jupyter.org/, consulté le 6 mai 2021

[17] Librairie Scikit, https://scikit-learn.org/stable/, consulté le 6 mai 2021

[18] Playground tensorflow, https://playground.tensorflow.org/, consulté le 6 mai 2021

[20] Debs, N., Rasti, P., Victor, L., Cho, T. H., Frindel, C., & Rousseau, D. (2020). Simulated perfusion MRI data to boost training of convolutional neural networks for lesion fate prediction in acute stroke. Computers in Biology and Medicine, 116, 103579.

[21] Jourdan, T., Boutet, A., & Frindel, C. (2018, November). Toward privacy in IoT mobile devices for activity recognition. In Proceedings of the 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (pp. 155-165).

[22] Fiche ECTS du module d'éthique de l'INSA Lyon, http://planete.insa-lyon.fr/scolpeda/f/ects?id=39503&_lang=fr, consulté le 6 mai 2021

[23] Atlantico, article du 21 septembre 2019, https://www.atlantico.fr/decryptage/3579561/ces-algorithmes-dont-on-ne-sait-rien-alors-qu-ils-regissent-nos-vies-amazon-google-facebook-daniel-le-metayer, consulté le 6 mai 2021

[24] Jourdan, T., Boutet, A., & Frindel, C. (2019). Vers la protection de la vie privée dans les objets connectés pour la reconnaissance d'activité en santé. Revue des Sciences et Technologies de l'Information-Série TSI: Technique et Science Informatiques.

[25] AskaBox, https://www.askabox.fr/, consulté le 6 mai 2021